



第一部分

Part One

初识数据

原始数据

用户编号	计量点编号	用户地址	行业 编号	时间	电量
yhbh	jldbh	yddz	xhylbdm	sjsj	dl
0800840011621380	08008400116213800001	广东省广州市...	IA0100	2017-01-01	72.60
0800150047281260	08001500472812600001	广州市番禺区...	IA0100	2017-01-01	76.00
0800840054270580	08008400542705800001	从化市鳌头镇...	IA0100	2017-01-01	73.20
0800830012924670	08008300129246700001	增城市中新镇...	IA0100	2017-01-01	79.77
0800140019640170	08001400196401700001	花都区花东镇...	IA0100	2017-01-01	76.50
0800840045312240	08008400453122400001	从化市良口镇...	IA0100	2017-01-01	79.64
0800830014194260	08008300141942600001	增城市朱村街...	IA0100	2017-01-01	79.61
0800150042859260	08001500428592600001	榄核镇雁沙村...	IA0100	2017-01-01	79.50
0800120037069980	08001200370699800002	九佛庚下九佛...	IA0100	2017-01-01	76.15
0800840011482390	08008400114823900001	从化江埔街上...	IA0100	2017-01-01	73.80

- 原始数据共 18,678,113条
- 时间跨度共709天(2017.0101-2018.12.10)
- 每天19,222—29,304条数据
- 共52,179用户;110个行业;

数据清洗

一、数据错误

- 重复记录 962条
- 3名用户缺少地址
- 1,929,395条数据存在0值记录

二、数据缺失

- 用户具有非0记录的天数: [1, 703]
- 多数用户时间不连续

三、数据不均衡

- 行业规模不均衡
- 单个用户属于单/多个行业
- 单个用户有单/多个计量点



第二部分

Part Two

相关性分析

相关性度量

01

互信息(Mutual Information)

定义了变量间相互依赖性的程度(对称)

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

02

转移熵(Transfer entropy)

定义了两个随机过程的信息传递量(不对称)

$$T_{X \rightarrow Y} = H(Y_t | Y_{t-1:t-L}) - H(Y_t | Y_{t-1:t-L}, X_{t-1:t-L})$$

03

皮尔森相关系数(Pearson Correlation Coefficient)

定义了变量间的直接相关性(对称)

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

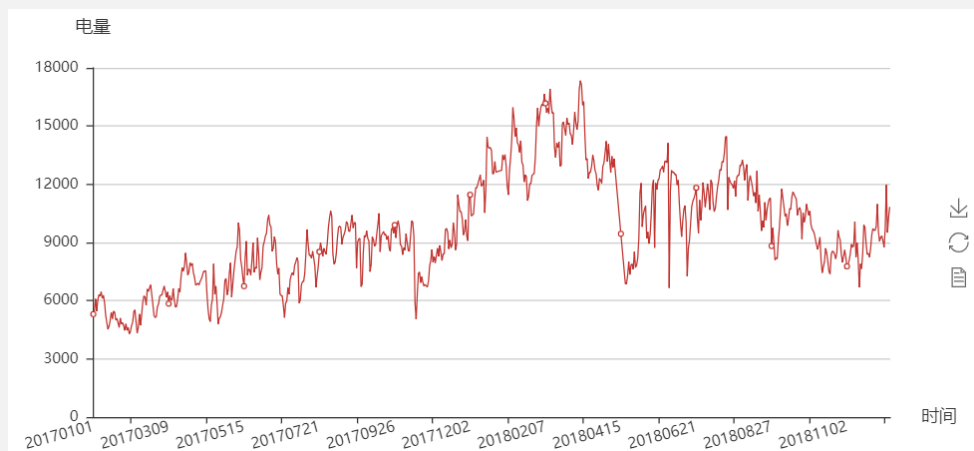
计算准备

一、数据过滤

- 过滤记录天数 ≤ 200 天的用户，并去重
- 过滤数据中存在用电量为0的记录
- 取各行业中日均用电量前五的用户，共 $105 \times 5 = 490$ 用户
- 其中大部分用户有大于400/709的用电量记录

二、数据补全

- 时间序列对齐，并通过插值补全某些日期缺少电量记录的数据



计算结果

一、相关性分析

- 通过线性插值补全计算的MI分布(左图): $[7.98 \times 10^{-6}, 2.27]$
- 通过线性插值补全计算的TE分布(右图): $[0, 0.247]$

$(-0.00226, 0.227]$	136065
$(0.227, 0.454]$	82240
$(0.454, 0.681]$	16802
$(0.681, 0.908]$	3989
$(0.908, 1.135]$	512
$(1.135, 1.362]$	32
$(1.362, 1.589]$	20
$(1.589, 1.816]$	69
$(1.816, 2.043]$	185
$(2.043, 2.27]$	186

MI

$(-0.000247, 0.0247]$	15524
$(0.0247, 0.0494]$	19018
$(0.0494, 0.0741]$	57537
$(0.0741, 0.0989]$	54686
$(0.0989, 0.124]$	44556
$(0.124, 0.148]$	31838
$(0.148, 0.173]$	13605
$(0.173, 0.198]$	2706
$(0.198, 0.222]$	566
$(0.222, 0.247]$	64

TE

二、数据异常

- 个别用户存在某些事件用电量异常高/低的情况
- 如用户'0800150005485840', 687天中686天用电量范围为 $[11.52, 67.87]$, 2017.08.18当天用电44143.03



第三部分

Part Three

研究问题

研究问题

(暂定)

一、问题概述

➤ 行业用电量相关性分析

二、研究内容

➤ 各行业两两间用电量的总相关性：

- (1) 行业间用电量相关性
- (2) 行业内用电量相关性

➤ 时序相关性：

- (1) 两用户的用电量相关性是否随时间变化

➤ 空间相关性：

- (1) 地理位置相近的用户相关性是否更强
- (2) 处于特殊区域的用户是否会对其他用户产生更大影响
- (3) 地理位置的周边设施分布是否对相关性有影响

研究方法

(暂定)

一、相关性计算

- 通过划分时间窗的方式，计算用户两两间的相关性

二、动态图构建

➤ 静态图构建：

- (1) 节点表示用户，边表示具有连接的两个用户具有强相关性
- (2) 一张静态图表示一个时间窗范围内，用户用电量的相关性关系
- (3) 必要时可以对节点聚合，将一个行业的用户聚合为一个节点

➤ 动态图构建：

- (1) 将一系列静态图按照时间顺序排列/连接
- (2) 通过行业、地理位置关系两种信息对图布局增加限制

研究方法

(暂定)

三、可视分析

- 各行业用电量统计（统计图表）
- 各行业用电量相关性总览（关系矩阵等）
- 动态图总览与分析（图可视化）
- 用户地理位置关系分析（地图）



第四部分

Part Four

任务安排

人员安排

一、人员安排

任务	人员
项目经理	卢金璇、陈则衔
前端实现	费治军
后端实现	路文杰

二、进度安排

- 4.1-4.15 进一步明确研究方法
 - 完成系统前后端框架搭建
 - 尝试构建动态图网络
 - 前后端技术调研
- 4.16-4.30 完成前端基础模块构建
 - 同步实现前后端数据接口
- 5.1-5.15 完成项目

绿色：已完成

红色：未完成

人员安排

三、4.3-4.10任务安排

路文杰：

- (1) 调研并学习JAVA后端技术
- (2) 进行后端搭建
- (3) 使用样例数据（490用户）计算划分时间窗的相关性变化
- (4) 后端需要传往前端的数据包括：用户/行业用电量、用户/行业两两间一段时间内的总相关性与相关性变化、动态图的拓扑结构等。

费治军：

- (1) 调研并学习前端技术，React、D3.js等
- (2) 进行前端框架搭建
- (3) 调研百度、高德地图api（用户数据目前只有中文地址，需要通过api转化为经纬度）
- (4) 绘制样例数据的相关性变化情况（折线图）

人员安排

三、4.10-4.17任务安排

总进度：

1.项目前后端搭建并整合完毕，使用git管理代码（费治军、路文杰）

后端(路文杰)：

(1) 计算动态图的拓扑结构

前端(费治军、卢金璇)：

(1) 绘制静态图结构（力引导）

(2) 动态图投影（参考reducing snapshots to points）

(3) 前后端接口整理

(4) 地图：根据上周调研情况安排工作

卢金璇：

(1) 阅读有关相关性分析的数据挖掘论文